Covid Twitter Discussions Analysis

Hong Van Pham, Yujia Luo, Xiyue Zhang
McGill University
van.pham2@mail.mcgill.ca

Introduction

The purpose of this report is to provide insights into recent Covid-19 English conversations on Twitter. More specifically, we developed topics discussed surrounding the current Covid situation based on 1000 tweets collected in a 72-hour window. We then conduct keyword analysis to characterize these topics and analyze the sentiment of the public towards them, with an emphasis on vaccine hesitancy. Twitter data was collected from Twitter Web API, queried by keywords relevant to the current Covid situation and vaccines, each data point was selected by the search query by id until the desired number of tweets were collected. No data aside from publishing time, tweet ID, and tweet content were collected.

To characterize the topics, we calculated and found 10 words of the highest tf-idf score for each. We then interpret the results with our understanding of the current Covid situation to make conclusions on the overall concerns and interest of the public of each topic. We also annotated each tweet for sentiment: positive, negative, and neutral. Due to time constraints, we could only perform single-annotation; however, we still managed to gather substantial data to gain insights on the public's overall sentiment regarding Covid-19 topics. The topic engagement was calculated using the percentage of each topic relative to the dataset to understand which topic was of most concern.

Given the limitation of time and resources, there are certain shortcomings in our method that influenced the results. During the data gathering process, due to the nature of the search query, tweets were not selected at random but by consecutive ID in the 3 days window. This could potentially create discussion "black holes" where if there is breaking news during this period, discussions would surround the specific topic heavily and skew the engagement and sentiment result. Eliminating duplicate tweets and retweets already limits the effect of these events. Our data collection will then provide an accurate understanding of very recent Twitter discussion, however, might be less accurate when making conclusions on the overall discussions surrounding Covid of a larger time window. Also, since Twitter is predominantly populated by users from the US and Canada, although our objective is to analyze English conversations surrounding the pandemic, our result will be heavily influenced by these countries' perspectives and sentiments. Twitter API policy also made it very difficult to filter tweets by location, and given the time constraints, it wasn't possible for this project.

We developed the 5 most discussed topics from the dataset, which are: attitude towards Covid, attitude towards vaccine, mandate, and policy, Covid cases, and Covid impact on the society and economy. Of the 1000 tweets, 30.0% were about vaccines, 27.7% on attitude towards Covid, 16.9% on mandate and policy. Covid's impact on the economy and society took up 14.8%, and 10.6% were on Covid cases. This result indicates that the discussions on Twitter mostly surround vaccination and that people mostly use Twitter to voice their attitude towards Covid, moreover, from the sentiment analysis, we recorded that the topic of vaccine and vaccination is the most divisive, with an almost equal number of tweets against and advocate for vaccines. This means although many are hesitant about getting vaccines or are against vaccination, there are also many who voice their support and encourage others to get the vaccine. Through the process of manual annotation, we can clearly see a difference between tweets containing vaccine attitudes and other tweets, with tweets about vaccinations being much more opinionated and passionate. The top keywords for this topic surround vaccination and shots. The only vaccine brand name appearing in this list is Pfizer, indicating it is the most notable and discussed brand of the Covid vaccine. The keywords "dose" and "booster" signify a discussion surrounding getting the recommended booster shots or the possibility of future, recurring Covid booster doses, which stems from the US government's recent authorization and recommendation for a booster Pfizer vaccine dose.

Our analysis also showed that the new Covid variant Omicron is discussed extensively, appearing in the list for "attitude towards Covid" and "Covid cases" top keywords. This is a result of the new cases of Omicron infections across the US and Canada are causing a lot of fear and anxiety in the public. News and tweets about new Omicron cases are widely shared, and many state their concern about the new variant as well as the possibility of another lockdown. It is also interesting how, in the "attitude towards Covid" topic, most of the positive sentiment stems from the belief that the Omicron variant may kill off other variants or may signify that the Covid virus is evolving to become a mild disease.

More detailed results and discussions will be presented below.

Data

We used Tweepy, a Python library, to access Twitter API and collect English tweets within a 72-hour window (from 00:00 27/11/2021 to 00:00 30/11/2021). The initial data set consists of 1094 tweets after the removal of exact duplicates using a simple Python function to compare tweets body. Manually going through the data set, we removed unrelated tweets and finalized the data set with 1000 tweets.

Replies and retweets were not collected for several reasons. Firstly, Twitter API v2 depreciated the "extended" option where one can access the full text of retweets when gathering the data. Collected retweets are cut short and often don't provide enough information in order to be analyzed. Secondly, when comparing data sets with and without replies and retweets, both using the same search queries, we noticed that the overall topic engagement and sentiments of the data set with retweets and replies significantly skewed to a particular topic or sentiment. This is due to the fact that Twitter API search is not exhaustive and might produce a bias in collecting tweets and retweets, for example, if out of 1000 tweets collected, retweets of one specific tweet take up 100 data points, the topic engagement and sentiment will likely be skewed depends on which tweet's retweets are chosen by the Twitter API. This will result in a misleading conclusion on the discussion around Covid. To minimize such an effect, retweets and replies were not collected so that the dataset more closely resemble the overall discussions around Covid and Covid vaccine on Twitter.

In order to collect relevant tweets, we used the following keywords: "covid", "vaccine", "vaccination", "pandemic", and covid vaccines brand names approved by the US and Canadian governments, "astra-zeneca", "pfizer", "moderna", and "johnson-and-johnson". We also included these words but as a hashtag, (e.g., "#covid" or "#vaccine") and a variation of covid-vaccine brand names (e.g., "astrazeneca", "johnsonandjohnson", "johnson&johnson", "janssen", "astra AND zeneca", …). Tweets collected will contain at least one of these keywords (in the case of "astra AND zeneca", tweet must include both "astra" and "zeneca").

We decided to only include brand names of vaccines that are approved by the Canadian and US governments because we are interested in discussions around vaccine sentiment in general, not about a specific vaccine brand. Looking at the same dataset with other vaccines such as "Sinovac" or "Sputnik" results in a significant amount of vaccine hesitancy tweets targeting these specific vaccines instead of the Covid vaccine in general. This might be due to factors such as prejudices, nationalism and other reasons which, although are important and interesting aspects of the current geopolitical landscape, are beyond the scope of this project.

Methods

In order to analyze the topics discussed and the theme and sentiment surrounding each topic, we developed 5 most prevalent topics and analyzed the most relevant keywords in each. We later use these words, our sentiment-annotation and understanding of the current Covid situation to conduct analysis and provide insights on topic engagement, sentiment surrounding each topic, and themes surrounding each topic.

We implemented the same approach as in assignment 8 to calculate the tf-idf score, namely:

- tf (w, t) = the number of times the word w occurs in the topic t
- idf (w, d) = log [(total number of topics) / (number of topics that have the word w)]
- => tf-df (w, t, d) =tf (w, t) * idf (w, d)

First, we manually annotated the data set for 5 topics and for sentiment (positive, neutral, negative). We then used Python to create a dictionary of all words appearing in each topic, removing words that occur less than 5 times throughout the data set for efficiency. We then calculate the top 10 words with the highest tf-idf score for each topic. From these 10 most discussed words, we can infer the most discussed topics within the 3-day window for each topic.

Aside from removing stop words and web links, we also removed contractions of stop words such as "gonna" and "imma", and typos, which commonly occur in the "attitude towards covid" topic. We also ignored words if their variations are already in the top 10 words. For example, "vaccines" and "vaccine" or "gov" and "government", are considered pairs of duplicates, while "vaccine" and "vaccinated" and "vaccination" are not. This elimination gives us a broader picture on the themes surrounding each topic, while still ensuring the integrity of the top keywords collected. The elimination will also help the analysis be more

straightforward and provide specific insights into the topic discussion.

We analyzed the engagement of each topic by calculating what percentage of the whole data set they take up. This will indicate which Covid-related topic is most relevant in recent conversations and highly reflective of what is most concerned by the public within the 3-day window.

We performed topic characterization by inferring the top 10 keywords using their tf-idf scores. From these keywords, we can understand what the most relevant discussions of each topic are, which subjects are of great concern during the Covid pandemic and combine the analysis with our own understanding of Covid situation to give insights into the public's sentiment and concerns.

Results

We used the first 200 tweets from the dataset for open coding and developing our topics. First, we seek if there is an existing typology to use. Fortunately, there are many scholars who did research on tweets' topic, trends and sentiments during the pandemic. Based on these theses, we found that covid-related tweets are most commonly vaccination, mandates, complaints and the pandemic's impacts, which is consistent with the first 200 tweets of our dataset. For this project, we are especially concerned about the response to the pandemic and vaccination. Thus, we selected the first two topics: (1) attitudes towards vaccination and (2) attitudes towards covid. After multiple attempts to categorize the 200 tweets, we decided the other 3 topics which are: (3) covid cases, (4) covid impact, (5) mandate/policy. Their definitions are presented in the table below. We also included some examples of tweets that fall into these categories for clarification.

Topic	Definition	Inclusion example
Attitudes towards vaccination	Tweets that are predominantly about public attitudes towards vaccination or any fact/news that could influence the attitudes towards vaccination	"Anti-vax Wayne County Republican is in ICU with COVID-19"
Attitudes towards covid	Tweets that express personal attitudes towards covid, usually are complaints or express fear about the pandemic, or how covid impacted their personal life.	"Aren't you tired of this pandemic? Really bone tired?"
Covid cases	Tweets that are case reports, either personal or by news sources, or tweets predicting future Covid infection rate and count.	"OMICRON INBOUND: It is only a "matter of time" that the COVID-19 variant Omicron will arrive in the United States and, eventually, Kansas."
Covid impact	Tweets on Covid's impact on the economy or the society in general.	"A sad, but not unexpected, development. Too bad, as I had a family member slated to compete." "World University Winter Games canceled due to COVID-19 variant."
Mandate/Policy	Tweets that are predominantly about mandates and policies for covid prevention, like lockdown or vaccine mandate, usually made by the government, companies or schools.	"BREAKING REPORT: Federal judge SUSPENDS Biden's COVID VACCINE MANDATE for health workers in 10 states."

Table 1: Topics and their definitions with examples.

Due to time constraint, single annotation was used instead of double annotation. Thus, the annotating process must follow the most-related rule. For instance, the inclusion example of Mandate/Policy in the table above also mentioned vaccination. However, it is more relevant to the policy and doesn't express the attitude towards vaccination, so the tweet falls into the Mandate/Policy topic.

After adjusting the topics definition and boundaries to fit the first 200 tweets, we decided to proceed with the above table and this typology is applied to the whole dataset. We manually went through the remaining 800 tweets to determine which topic each falls into. We also code the sentiment

for each tweet, by positive, negative, and neutral. The relative engagement with sentiments in each topic are displayed in this chart below.

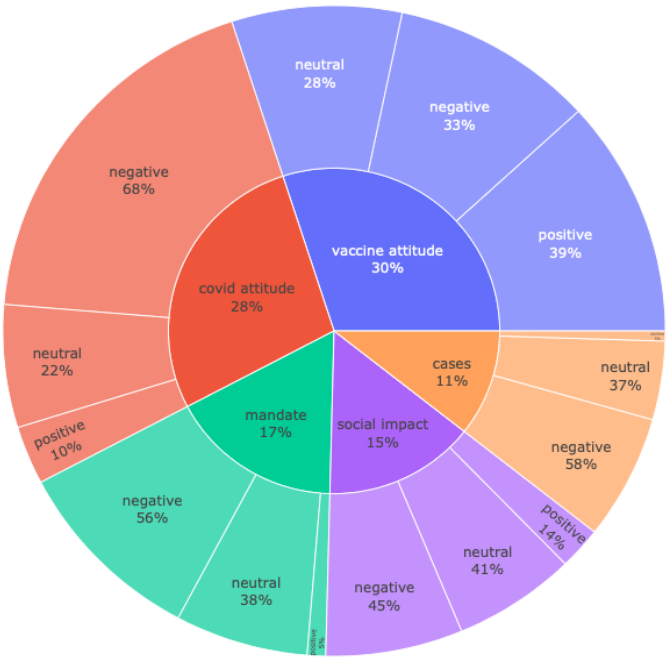

Chart 1: Relative engagement by topic and their respective sentiment ratio

Of the 1000 tweets, 30.0% were about vaccines, 27.7% on attitude towards Covid, 16.9% on mandate and policy. Covid's impact on the economy and society took up 14.8%, and 10.6% were on Covid cases. Their relative sentiments are also presented above. Notably, "attitudes towards vaccines" (denoted "vaccine attitude") is the most engaged topic as well as the most divisive topic, with an almost equal number of tweets with negative and positive sentiments. The topic on "attitudes towards Covid" (denoted "Covid attitude") is the 2nd most engaged topic. It received the largest percent of negative discussions, as expected. Overall, out of the 1000 tweets, 51% were of negative sentiment, 30.8% were of neutral sentiment and 18.2% were of positive sentiment.

Words with highest tf-idf scores in each topic are calculated and presented below. In our research, we assume that these words are most frequently used and they can present the discussions of these topics.

Topic	Keywords
Attitude towards vaccine	pfizer, shot, vaccine, booster, adults, dose, authorization, vaxxed, moderna, fda
Attitudes towards Covid	sick, real, believe, variant, pandemic, tested, omicron, please, tired, break
Covid cases	detected, analytics, alberta, canada, wave, potential, omicron, variant, deaths, active
Covid impact	investment, labor, fundamentally, reshaping, workforce, wealth, investors, reduce, impact, pandemic
Mandate/policy	mandate, judge,vaccine, federal, biden, workers, delay, blocks, court, pass

Table 2: Top 10 keywords of each topic ranked by tf-idf score, in decreasing order.

Discussion

As for the vaccine topic, it is not surprising that pfizer is the most frequent word, which is the most widely discussed brand name vaccine and vaccine company, followed by another vaccine brand, moderna. Additionally, the keyword "fda", "booster", and "adults" signifies that there are many discussions surrounding the recommended booster shots or the possibility of recurring Covid booster doses, which stems from the US government's recent authorization and recommendation for adults to get a booster vaccine dose.

The sentiment ratio for vaccines is 39% positive, 33% negative, 28% are neutral, indicating that this is a very divisive topic. This is due to the fact that many people use Twitter to voice their opinion on vaccine effectiveness and advocate other people to follow their belief as well. The recent news about the booster dose and the Omicron variant surely raised many questions on the Covid vaccine potency, while many try to encourage others to get the booster shot to reduce the new variant's growth and spread.

Further, we wanted to understand the current discussion involving vaccine hesitancy, since a substantial proportion of tweets are negative indicating that there are many who are skeptical or against the vaccines. To gain more detailed insight for this, we rerun the code for tf-idf scores, this time limiting to tweets on the topic of vaccines with negative sentiment. We get a few different words, which are "effective", "wrong", and "dead" in addition to the original keywords.

We see that the negative sentiment towards Covid vaccine mostly stems from its apparent inability to stop the spread of Covid, it's supposed incapability to reduce Covid related deaths, and the government's ineptitude (or unwillingness) to find a cure for Covid. Besides, in the mandate topic, the word "vaccine" appears frequently in negative sentiment tweets. These results indicate that doubts about the effectiveness of vaccines and aversion to mandatory vaccine policies may have caused vaccine hesitancy and resistance.

The results for attitudes towards covid are roughly consistent with our expectations. The relevant topical keywords are mostly of negative or desperate emotions. For instance, the word "sick" implies 2 themes within the attitude Covid. The more straightforward approach is referring to how people are getting physically ill, with a possibility of it being covid infections. The other approach, which is supported by the keyword "tired" refers to people getting "sick and tired" of the current Covid situation, frustrated by the anxiety and fear of Covid's impact on their health and other aspects in life. An interesting finding here is we assumed no tweet would have positive sentiment. 68% of tweets in this topic are negative and 22% are neutral. Unexpectedly, there are 10% positive tweets. Some people said, the covid changed their life for the better. And some expressed their optimism for the future of the pandemic. These two categories count for 10%.

The negative sentiment ratio is 56% for the mandate/policy topic. When annotating, we also noticed that when people talk about the regulations made by the government, they are rarely supportive. Keywords such as "federal", "biden", "court" refer to the fact that most discussions in this topic are from the United States. And people most care about lockdowns, vaccine passes and vaccine mandates for workers, which are inferred from "blocks", "pass" and "workers".

The covid impact might be the broadest topic among the five. The keywords indicate the financial sector and industry are influenced most by covid. And the impact of covid is destructive and long lasting. "Investment", "wealth", "investors" refer to the concern of Covid's impact over the economy, specifically in the financial sector where Covid's impact is most uncertain, and while the global economy is slowly returning to the pre-Covid trend, investments are of crucial importance on which industry will grow after the pandemic. "Labor" and" workforce" implies that discussions also surround covid impact on the workers. During the pandemic, the workers were not only affected by the change in working conditions due to social distancing regulations, but many were also laid off due to mass downsizing, or whose choice of work is limited due to vaccine mandates. The nature of the workforce after the economy returns to the pre-covid trend and social distancing ends is shown to be of great interest according to the analysis.

The keywords in topic covid cases display a trending of covid spread in north America, which clearly suggest the predominant concern over the new Omicron variant cases detected. The Covid variant was recently discovered in the week leading up to the data collection, so news about new cases is of great concern and are updated frequently, as the result shows. "Canada" and "Alberta" might suggest in the period we selected; the situation of the pandemic is serious over Canada.

Conclusively, the pandemic has a serious impact on people's lives, but the vaccine policy and the government's pandemic prevention measures have been challenged by public opinion. The emergence of new variants makes the pandemic more uncertain. If we want to increase the vaccination rate, the government and public health departments may need more efforts to persuade vaccine hesitant individuals. The analysis also suggests that the current mandates and policies of government agencies and representatives are receiving widely negative responses from the public, and further expansion on vaccine mandates might cause more vaccine hesitancy and overall dissatisfaction and distrust towards the government.

Despite our attention to details, due to time constraints, there are several factors that could potentially be improved to produce more accurate results. When selecting the tweets, we only used the keywords as the content filter, which may cause some information bias. Analyzing datasets of different queries or implementing more advanced algorithms to choose tweets might be needed for a comprehensive selection. We can also implement machine learning to help develop and verify our topics. The 3-day length chosen for data gathering might also cause some problems. A short window such as 72 hours is good for analysis of very recent dates but is generally insufficient to analyze the overall Covid discussion of a longer time window. We also couldn't filter tweets by geolocations or countries. Even though the project's objective was to analyze English tweets, most discussions selected are from the US and Canada due to how Twitter is populated by users from these two countries. This gives a very American, Canadian-centric dataset, but we still couldn't rigorously narrow the analysis target to these two countries since these tweets can also be from or about other English-speaking countries. Another way to improve and expand our analysis is through double annotation, which will render the results to be more informative and accurate.

There are many directions to expand this to a future project. One of such projects is to conduct a vaccine hesitancy analysis on other Covid vaccine brand names that are not approved by the US and Canadian government. We can also conduct analysis on other social media platforms like Reddit or Facebook to have a broader picture. Twitter users might fall into a category which might not give an accurate representation of overall Covid discussions on social media. These expansions can give a more informative result and

produce a more representative picture on the different discussions and sentiment surrounding the pandemic.

Group Member contributions

Hong Van Pham was in charge of data collection and keyword analysis. Xiyue Zhang was responsible for topic development and engagement analysis. Yujia Luo was tasked with topic annotation. The discussions were written by the group collectively.

References

Tweepy. https://www.tweepy.org/. Accessed: 2021-11-29.

Kwok S, Vadde S, Wang G. 2021.
Tweet Topics and Sentiments Relating to COVID-19 Vaccination Among Australian Twitter Users: Machine Learning Analysis. https://www.jmir.org/2021/5/e26953
DOI: 10.2196/26953. Accessed: 2021-11-30.

Chandrasekaran R, Mehta V, Valkunde T, Moustakas E. 2021. Topics, Trends, and Sentiments of Tweets About the COVID-19 Pandemic: Temporal Infoveillance Study. https://www.jmir.org/2020/10/e22624
DOI: 10.2196/22624. Accessed: 2021-11-30.